DEPARTMENT OF MATHEMATICS
UNIVERSITY OF NIJMEGEN The Netherlands

ON GENERALIZED CONJUGATE GRADIENT TYPE METHODS FOR THE ITERATIVE SOLUTION OF NONSYMMETRIC AND/OR INDEFINITE SYSTEMS OF EQUATIONS; GENERAL CONVERGENCE PROPERTIES

Owe Axelsson

Report No. 9903 (January 1999)

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF NIJMEGEN
Toernooiveld
6525 ED Nijmegen
The Netherlands

On generalized conjugate gradient type methods for the iterative solution of nonsymmetric and/or indefinite systems of equations; general convergence properties

Owe Axelsson
Department of Mathematics
University of Nijmegen
Nijmegen, The Netherlands

Dedicated to David M. Young on the occasion of his 75th birthday.

Abstract

The behaviour of iterative solution methods for linear systems of algebraic equations, in general nonsymmetric and/or indefinite, is considered. The methods analysed are generalized conjugate gradient methods of minimal residual or orthogonal residual type using a Krylov set of vectors, defined by the preconditioned matrix B, and/or other vectors.

A general convergence result showing convergence for any matrix whose field of values does not contain the origin, is given. The rate of convergence can be analysed using the spectrum, pseudo-eigenvalues, or more generally the field of values of B.

For severely ill-conditioned and/or strongly non-normal matrices convergence stagnation occurs but can be avoided using restart of the method with a properly chosen new preconditioner.

An iterative solution method is best characterized by being either a monotonically convergent minimization method or, more generally, just a projection method on a subspace of vectors. A relation in the form of peaks and plateaus between the convergence of pairs of methods from either class is shown.

Other issues discussed include: automatic truncation to a short length version, the use of normal equations and efficient implementation of the methods, with minimal complexity per iteration step.

The presentation is divided in two parts. Part one, i.e. the present paper, deals with the characterization and convergence, including stagnation of the method. Part two, to appear in a separate paper deals with the remaining topics.

Keywords: Iterative solution, generalized conjugate gradient, characterization, stagnation, variable preconditioners, pseudo-eigenvalues, field of values.

1 Introduction

Consider the iterative solution of a linear system $Ax = b$, where A is an $n \times n$ matrix, $x \in \mathbb{C}^n$, $b \in \mathbb{C}^n$. Our major concern here is solving such systems where A is non-hermitian and/or indefinite. The classical conjugate gradient method of Hestenes and Stiefel [18] was derived as a method to solve symmetric (or hermitian) and positive definite systems, using iteratively computed approximations from a Krylov vector subspace, $V_{k-1}(d^0, A) = \text{span}\{d^0, Ad^0, \ldots, A^{k-1}d^0\}$, where d^0 is a given vector, such as $d^0 = Ax^0 - b$, for some initial vector x^0.

Over the years since then, a number of generalized conjugate gradient methods have been proposed to solve more general, non-hermitian and/or indefinite systems. Some of the methods can form the iterations on more general vector subspaces than the Krylov space. Various attempts have been made to characterize the methods, see [2], [10], [17], [22], [19].

These characterizations have been based on the inner product used (normally defined by an Hermitian positive definite matrix, H) and the matrices A, A^T, B, B^T, where $B = C^{-1}A$ and C is a preconditioner, which may be involved in the method.

However, a more natural way to characterize such iterative methods seems to be by their fulfillment of one of the major properties: minimization, orthogonalization based on projection, or orthogonalization without projection. Let W be a vector space with inner product $(\cdot, \cdot)$ in W and let V, U be subspaces of W.

By projection of a vector r^0 onto a subspace V we mean a vector $g \in V$ such that

$$(g, v) = (r^0, v) \qquad \text{for all } v \in U,$$

or,

$$(g - r^0, v) = 0 \qquad \text{for all } v \in U.$$

We shall consider cases where V and U are finite dimensional of the same dimension in which case g is unique. For the standard inner product, such a projection is also called the l_2-projection.

This leads to the following characterizations:

(i) methods based on minimization of the residual $r^k = b - Ax^k$, or the iteration error, $x - x^k$ in some norm over a vector subspace, where x^k is the current approximation to the vector x to be computed. This corresponds to a projection, called Ritz projection, where $U = V$.

(ii) methods based on making the residual or error orthogonal w.r.t. another vector space U than V. This projection is called Galerkin projection.

(iii) methods based on making the residual or error orthogonal w.r.t. a generalized inner product, i.e., one which is defined by a symmetric but not necessarily positive definite, but nonsingular matrix. A typical inner product is here given by

$$(\mathbf{u}, \mathbf{v}) = \mathbf{u} \begin{pmatrix} 0 & A \\ A^* & 0 \end{pmatrix} \mathbf{v}.$$

Note that this inner product does not define a norm.

Examples of familiar methods within each class are

(i) ORTHOMIN (Vinsome, 1976), GCG-LS (Axelsson, 1980, 1987), GMRES (Saad and Schultz, 1986).

(ii) ORTHORES (Young and Jea, 1980), GCG-OR (Axelsson and Makorov, 1995), (Axelsson, 1994)

(iii) BCG (Fletcher, 1976; Jea and Young, 1983 for such a presentation of the biorthogonal Lanczos method)

Note that all methods in the same class which are based on the same inner product and subspaces give the same approximation in exact arithmetic. It is only their implementations which may differ.

For further remarks regarding the characterization of generalized conjugate gradient methods, see [28].

In the paper we will frequently refer to the GMRES (generalized minimal residual) method [23] which has become a popular method. We recall that it is a Krylov subspace method in which first the basis vectors for $V_{k-1}(d^0, A)$ are calculated via an orthogonalization process, referred to as Arnoldi process, see [1].

The vector x^k is then computed at predetermined steps (k) such that

$$\|b - Ax^k\|_2 = \min_{x \in x^0 \oplus V_{k-1}(d^0, A)} \|b - Ax\|_2.$$

Computing the next basis vector requires only one matrix vector multiplication with A (and a preconditioning step in the preconditioned version of the method).

In general, there is no a priori information available to tell when the residual $r^k = Ax^k - b$ is sufficiently small, so one may do more steps than required for the wanted accuracy. All basis vectors have to be stored and the number of vector operations (inner products and vector additions) increases linearly with k as, in general, no short recursion is available. Therefore for large values of k, the method becomes less efficient and the GMRES method is normally restarted after every k_0 iterations for some k_0, typically of the order 10-20, using the current iterate as a new initial vector.

The present paper discusses in a survey form three fundamental issues related to the behaviour of such iteration methods, namely

(i) convergence properties, such as best approximation properties, stagnation and avoidance of stagnation.

(ii) relations in the form of peaks and plateaus between certain pairs of methods taken from the different classes

(iii) implementational aspects.

The paper is divided in two parts of which this is the first part.

It is organized as follows. In Section 2 the generalized conjugate gradient method is presented. Its major convergence properties are shown in Section 3 and some

concluding remarks are found in Section 4.
The second part discusses the importance of using non-Krylov subspaces of vectors to avoid stagnation of the convergence and shows a relation in the form of peaks and plateaus between pairs of methods from the Galerkin and Ritz projection classes of methods. Some implementational aspects of the GCG type methods are discussed and a comparison with the GMRES method is made. In the final section some concluding remarks on automatic truncation to a short length version and convergence properties of normal equation approaches are found.

2 The GCG-MR and GCG-OR methods with variable preconditioners

Let $(u, v) = u^* H v$, where H is h.p.d., define an inner product, let $\|u\| = (u, u)^{\frac{1}{2}}$ and let

$$a(u, v) \equiv (Au, Av) \text{ ,for the GCG-MR method} \tag{2.1}$$

$$a(u, v) \equiv (Au, v) \text{ ,for the GCG-OR method} \tag{2.2}$$

To solve the linear system $Ax = b$, where A is nonsingular, following [3], [4], [5], [8] we present now the generalized conjugate gradient-minimum residual (GCG-MR) and the generalized conjugate gradient-orthogonal residual (GCG-OR) methods in a form which accommodates the use of variable preconditioners.

The methods presented here use a long recursion for the update of the solution vector x^k instead of for the search direction vector d^k as in the other similar (and, in exact arithmetic equivalent) methods within the same class. As we shall see, thereby one must solve small sized least squares problems whereas the orthogonalization of the vectors $\{d^k\}$, required in the other methods, is avoided. The methods used here are therefore not, or less, influenced by the loss of exact orthogonality.

Given $s \geq 1$, where s is the truncation index, i.e., the maximum number of search directions vectors used at any stage, a sequence of nonsingular preconditioners C_k, $k \geq 0$ and an initial vector x^0, let

$$\begin{aligned} r^0 &= b - Ax^0 \ , \\ d^0 &= C_0^{-1} r^0 \ . \end{aligned}$$

At step k, the vector x^k in the sequence $\{x^k\}$ approximating the solution x, is defined by

$$x^k = x^{k-1} + \sum_{j=k-s_k}^{k-1} \alpha_j^{(k)} d^j \ , \qquad 1 \leq s_k \leq \min(k, s) \ , \tag{2.3}$$

where $\{d^j\}$ are search directions and $\{\alpha_j^{(k)}\}$ are parameters computed to do one of the following:

(a) minimize (r^k, r^k), where $r^k = b - Ax^k$ (in the GCG-MR method)

or

(b) make r^k orthogonal to the s_k previous search directions, i.e. make

$$(r^k, d^\ell) = 0 \ , \qquad k - s_k \leq \ell \leq k-1 \tag{2.4}$$

(in the GCG-OR method).

Relation (2.3) shows that

$$r^k = r^{k-1} - \sum_{k-s_k}^{k-1} \alpha_j^{(k)} A d^j \ , \tag{2.5}$$

which can be used as an alternative to computing the residuals by the defining relation $r^k = b - Ax^k$, if it is more cost efficient and if the propagations of round-off errors is on a satisfactory level, to prevent the recursively computed residuals from becoming too different from the true residuals.

By the minimization property in (a), the residual r^k is orthogonal to the subspace AV_{k-1}, where

$$V_{k-1} = \text{span}\{d^{k-s_k}, \ldots, d^{k-1}\} \ ,$$

because r^k is the difference between r^{k-1} and the projection of r^{k-1} onto AV_{k-1}. Hence

$$(r^k, Av) = 0 \ , \quad \text{for all} \quad v \in V_{k-1}$$

or, by (2.1),

$$a(e^k, v) = 0 \ , \quad \text{for all} \quad v \in V_{k-1} \tag{2.6}$$

where

$$e^k = x - x^k = A^{-1} r^k \ , \tag{2.7}$$

is the iteration error. By (2.4) and (2.2), relation (2.6) holds for the GCG-OR method also.

The new search-direction vector can be defined by

$$d^k = C_k^{-1} r^k \ . \tag{2.8}$$

Alternatively, it can be defined by

$$d^k = C_k^{-1} r^k - \beta_k d^{k-1} \ , \tag{2.9}$$

where β_k is a scalar parameter computed to make

$$a(d^{k-1}, d^k) = 0 \ ,$$

that is,

$$\beta_k = \frac{a(d^{k-1}, C_k^{-1} r^k)}{a(d^{k-1}, d^{k-1})} \ . \tag{2.10}$$

As it turns out, the latter choice gives automatic truncation i.e., a vector recursion of length $s = 1$ if $C_k = C$ is fixed and $C^{-1}A$ is an H-normal matrix (see [12], [20],

[4], for instance). If $C_k = C$, the expression in (2.10) can be simplified, similar as has been done in [8]. Due to a possible division by zero in (2.10) in the case (b), one should use (2.9) only when A is positive definite w.r.t. the inner product used.

It is also possible to use a longer recursion than in (2.9) and orthogonalize the vectors, possibly with respect to some other inner product. In particular, to avoid the long recursion one can generate orthogonal vectors such as in the BCG method, which requires only a three term recursion. Numerical tests have shown that even if the least squares minimization is done with just a few vectors, the GCG method can offer stabilization of the erratic convergence behavior seen in the pure BCG method. An example of such a method is the BCG-Stab method (see [26]) where just a one-dimensional minimization is done. The GCG method offers a general approach for such stabilizations.

Relations (2.6), (2.7) and (2.5) show that

$$\sum_{j=k-s_k}^{k-1} \alpha_j^{(k)} a(d^j, d^\ell) = a(e^{k-1}, d^\ell) \ , \qquad \ell = k-1, \ldots, k-s_k \ ,$$

or

$$\Lambda^{(k)} \alpha^{(k)} = \boldsymbol{\gamma}^{(k)} \tag{2.11}$$

where

$$\Lambda_{\ell,j}^{(k)} = a(d^j, d^\ell), (\alpha^{(k)})_j = \alpha_j^{(k)} \ , (\boldsymbol{\gamma}^{(k)})_\ell = \gamma_\ell^{(k)}$$

and

$$(\boldsymbol{\gamma}^{(k)})_{k-1} = \begin{cases} (r^{k-1}, Ad^{k-1}), & \text{in case (a)} \\ (r^{k-1}, d^{k-1}) & \text{in case (b)}, \end{cases} ,$$
$$\gamma_\ell^{(k)} = 0 \ , \qquad k - s_k \leq \ell \leq k-2.$$

Relation (2.11) is used to compute the coefficients in (2.3).

The matrix $\Lambda^{(k)}$, which has order $s_k \times s_k$ has the following properties:

Case (a): Here $a(d^j, d^\ell) = (Ad^j, Ad^\ell)$, so $\Lambda^{(k)}$ is symmetric. Furthermore, it is positive definite if and only if the vector set V_k is spanned by linearly independent vectors $\{d^j\}_{k-s_k}^{k-1}$. (Note here that we have assumed that A is nonsingular.) Due to the minimization property, it follows that this set is linearly dependent if and only if $r^{k-1} = 0$, i.e., the solution has already been found. Hence, the method does not suffer any breakdown.

Further, $\Lambda^{(k)}$ equals $\Lambda^{(k-1)}$ augmented with a row and column. This observation is important from a computational complexity point of view as only s_k inner products must be computed in all. In addition, at a truncation step, where $s_k \leq s_{k-1}$ (some of) the earlier row(s) and column(s) are deleted. Hence choosing the parameter s_k properly one can limit the total computational complexity. In practice, the matrix $\Lambda^{(k)}$ can become very ill-conditioned when k is large. This may cause unacceptably large rounding errors. One must therefore combine the method with restarts as is also normally done in the GMRES method.

Case (b): Here $a(d^j, d^\ell) = (Ad^j, d^\ell)$.

Without further assumptions, in this case $\Lambda^{(k)}$ may become singular, and the method will then suffer a breakdown. Assume now that we use a fixed preconditioner so that $C_k = C$, $k = 0, 1, \ldots$. Then it can be readily seen (see also [8]) that (2.6), (2.9) and (2.5) imply that

$$a(d^j, d^\ell) = 0 \ , \qquad 0 \le j \le \ell - 1$$

that is, the matrix $\Lambda^{(k)}$ is upper triangular. Further, the leading coefficient in (2.3) satisfies

$$\alpha_{k-1}^{(k-1)} = (r^{k-1}, r^{k-1})/a(d^{k-1}, d^{k-1}) \ .$$

Hence, as long as $a(d^{k-1}, d^{k-1}) \neq 0$ (which holds if A is positive definite w.r.t. the inner product) the method will not suffer any breakdown before the solution has been found. However, if $a(d^{k-1}, d^{k-1}) = 0$, $\Lambda^{(k)}$ is singular and the method may have breakdown.

3 Convergence behaviors of the GCG-MR and GCG-OR methods

Convergence of iterative methods is typically measured by the ratio $\|r^k\|/\|r^0\|$ or by the ratio $\|e^k\|/\|e^0\|$. However, normally the latter is not available. In this section a general convergence estimate based on the relative residuals is first presented followed by some specific estimates based on eigenvalues and on pseudo-eigenvalues.

3.1 General convergence estimates

To analyze the convergence of the methods consider first the full, untruncated method where $s_k = k$, and assume that $C_k = C$, $k \ge 0$. By (2.3) we have

$$x^k = x^0 + \sum_0^{k-1} \xi_j^{(k)} d^j$$

for some coefficients $\xi_j^{(k)}$.

In case (a), by construction we have

$$\begin{aligned} (r^k, r^k) &= a(x - x^k, x - x^k) \\ &= \min a(x - v, x - v) \ , \qquad v \in x^0 \oplus V^{k-1} \end{aligned} \tag{3.1}$$

since $v = x^0 + p_{k-1}(C^{-1}A)d^0$, where $p_{k-1}(\cdot)$ is a polynomial of degree $k-1$ and $p_{k-1}(0) = 0$. As has been shown in [4], [5], for instance, the optimal approximation property (3.1) can be used to give lower bounds of the rate of convergence using various polynomial approximation properties. As it turns out, the rate of convergence depends heavily on the distribution of eigenvalues of $C^{-1}A$. For the case where the eigenvalues are all located in one (or several) ellipse(s) in the complex plane, see [6] and the references therein, and the next subsection.

As shown in [4], the following general estimate holds for the rate of convergence of the GCG-MR method.

Theorem 3.1 *Consider the GCG-MR method (2.3), (2.11)). Let*

$$V_{k,s_k} = span\{d^{k-s_k}, \ldots, d^{k-1}, d^k\}.$$

Then successive residuals are related as

$$\min_{v \in BV_{k,s_k}} \|r^k - v\|^2 = (r^{k+1}, r^{k+1}) = (r^k, r^k) - \left(\frac{(r^k, Br^k)}{\min\limits_{w \in BV_{k-1,s_k-1}} \|Br^k - w\|} \right)^2$$

where $B = C^{-1}A$.

Remark 3.1 Theorem 3.1 implies *monotone convergence*, i.e. $\|r^{k+1}\| < \|r^k\|$, when $(r^k, Br^k) \neq 0$. Further, we have seen that $\Lambda^{(k)}$ is singular if and only if the search directions become linearly dependent, which can only occur when $r^k = 0$. If the field of values, $W(B) = \{(z, Bz); z \in \mathbb{C}^n, \|z\| = 1\}$ does not contain zero, then the hermitian part of B (w.r.t. the inner product defined by H, i.e., the hermitian part of $\widetilde{B} = H^{-\frac{1}{2}} B H^{\frac{1}{2}}$) is positive (or negative) definite, and $(r^k, Br^k) = 0$ if and only if $r^k = 0$, i.e., the solution has already been found. This follows since the real part of $W(B)$ is equal to the interval $[\lambda_{\min}(\frac{1}{2}(\widetilde{B} + \widetilde{B}^*)), \lambda_{\max}(\frac{1}{2}(\widetilde{B} + \widetilde{B}^*))]$. It can be shown that $W(B)$ does not contain zero if $W(B)$ lies in any open half-space $\{z; Re(e^{-i\theta} z) > 0\}$.
For $\min \|Br^k - w\|^2$ a similar expression as for $\|r^{k+1}\|^2$ holds which shows that the estimate in Theorem 3.1 is of a continued fraction type. (See [4] for further details.)

Remark 3.2 Letting $w = 0$ in Theorem 3.1, we obtain the upper bound

$$(r^{k+1}, r^{k+1}) \leq (r^k, r^k) - \left(\frac{(r^k, Br^k)}{\|Br^k\|} \right)^2.$$

This estimate is the same as for the steepest descent method (which corresponds to $s = 1$). If $H = I$, we have then

$$\|r^{k+1}\|^2 \leq \left[1 - \frac{(r^k, Br^k)}{(r^k, r^k)} \cdot \frac{(r^k, Br^k)}{(Br^k, Br^k)} \right] \|r^k\|^2.$$

Let θ_k be the acute angle between r^k and Br^k. Then

$$\cos\theta_k = \frac{|(r^k, Br^k)|}{\|r^k\| \, \|Br^k\|}$$

and the estimate takes the form

$$\|r^{k+1}\|^2 \leq (1 - \cos^2\theta_k) \|r^k\|^2$$

or

$$\|r^{k+1}\| \leq \sin\theta_k \|r^k\|.$$

For this and other convergence estimates involving "operator trigonometry" see [16]. Note that $\cos\theta_k$ can be used to tell when there is convergence stagnation.

Another bound is

$$\|r^{k+1}\|^2 \leq \left(1 - \min_r \frac{(r,Br)}{(r,r)} \min_s \frac{(s,B^{-1}s)}{(s,s)}\right) \|r^k\|^2, \tag{3.2}$$

that is, this upper bound involves the product of the smallest real parts of the field of values of B and of B^{-1}. In general, as shown in Theorem 3.1, the rate of convergence of the GCG-MR method is much faster than what the latter bounds predict.

Note also that even if A and C are hpd, the hermitian part of $B = C^{-1}A$ may be indefinite, so requiring $\lambda_{\min}(\frac{1}{2}(B+B^*))$ or $\lambda_{\min}\left(\frac{1}{2}\left(B^{-1}+B^{*^{-1}}\right)\right)$ to be positive may not be feasible in practice. A simple example illustrating this is

$$C^{-1}A = \begin{bmatrix} 1 & 0 \\ 0 & 0.1 \end{bmatrix} \begin{bmatrix} 1 & -3 \\ -3 & 10 \end{bmatrix} = \begin{bmatrix} 1 & -3 \\ -0.3 & 1 \end{bmatrix}$$

whose symmetric part is $\begin{bmatrix} 1 & -1.65 \\ -1.65 & 1 \end{bmatrix}$, i.e., indefinite.

The GCG-OR method

For method (b) we follow [5], [8], to show a similar bound as in (3.1), which holds if the bilinear form $a(\cdot,\cdot)$ is *positive definite*, that is, we assume that there exists a positive number ρ such that

$$a(u,u) \geq \rho(u,u)\ , \quad \text{for all} \quad u \in \mathbb{C}^n \tag{3.3}$$

It is readily seen that this holds if and only if the symmetric part of $\tilde{B} \equiv H^{-1/2}BH^{1/2}$ is positive definite, and then $\rho = \lambda_{\min}(\frac{1}{2}(\tilde{B}+\tilde{B}^*))$. Furthermore, we assume that $a(u,v)$ is *bounded*, i.e., there exists a constant K, $K \geq \rho$ such that

$$|a(u,v)| \leq K(u,u)^{1/2}(v,v)^{1/2}\ , \quad \text{for all} \quad v \in \mathbb{C}^n\ .$$

This holds with $K = \|\tilde{B}\| = \lambda_{\max}(\tilde{B}^*\tilde{B})^{1/2}$. Clearly $K \geq \rho$. If $\widetilde{B}$ is spd, then $\rho = \lambda_{\min}(\widetilde{B})$ and $K = \lambda_{\max}(\widetilde{B})$.

Using the orthogonality property (2.6),

$$a(e^k, v) = 0\ , \quad \text{for all} \quad v \in x^0 \oplus V_{k-1}$$

and the above bounds on $a(u,v)$, we find

$$\begin{aligned} a(e^k,e^k) &= a(e^k, x-v) \\ &\leq K(e^k,e^k)^{1/2}(x-v,x-v)^{1/2} \end{aligned}$$

for all $v \in x^0 \oplus V_{k-1}$. This and (3.3) show the quasioptimal (quasioptimal, because $K/\rho \geq 1$, in general) property,

$$\|e^k\| = (e^k, e^k)^{1/2} \leq \frac{K}{\rho} \min_{x \in x^0 \oplus V_{k-1}} (x - v, x - v)^{1/2} \tag{3.4}$$

and the average rate of convergence,

$$\|e^k\|^{1/k} \leq \left(\frac{K}{\rho}\right)^{1/k} \min_{x \in x^0 \oplus V_{k-1}} \|x - v\|^{1/k} \ .$$

Since $(K/\rho)^{1/k} \to 1$, $k \to \infty$, it follows that in exact arithmetic the average error becomes arbitrary close to the best approximation error. (Typical values could be $K/\rho = 10^4$ and $k = 32$, in which case $(K/\rho)^{\frac{1}{k}} \approx 1.35$.) Hence, the average convergence rate approaches that of method (a), but in a different norm.

3.2 Convergence estimates based on eigenvalues and on pseudo-eigenvalues

As has been shown in [4], [8], (3.1) and (3.4) can be used to estimate the rate of convergence of the GCG-MR and GCG-OR methods when $C_k = C$. The estimates are based on the fact that for any matrix B ($B = C^{-1}A$) there exists a nonsingular matrix S such that a Jordan decomposition

$$S^{-1}BS = J \ ,$$

of B holds, where J is block diagonal and each diagonal block J_i is itself either diagonal or a Jordan matrix, of order s_i of the form

$$J_i = \begin{bmatrix} \lambda_i & 1 & & 0 \\ & \ddots & \ddots & \\ & & & 1 \\ 0 & & & \lambda_i \end{bmatrix}$$

The number s_i is called the deficiency index, since $s_i - 1$ is the number of deficient eigenvalues corresponding to the single eigenvector of J_i. From the above it follows now readily that

$$\min_{v \in x^0 \oplus V_{k-1}} \|x - v\| \leq \|S\| \, \|S^{-1}\| \min_{p_k \in \pi_k^1} \|p_k(J)e^0\| \ ,$$

where π_k^1 denotes the set of polynomials of degree k, normalized at the origin.

The number of iterations required to make

$$\|e^k\| \leq \varepsilon \|e^0\| \tag{3.5}$$

for some $\varepsilon > 0$ can then be estimated as the smallest integer k, such that for some polynomial $p_k(J)$ of degree k we have

$$\min_{p_k \in \pi_k^1} \|p_k(J)e^0\| \leq \frac{\rho\varepsilon}{K\kappa(S)}\|e^0\| \tag{3.6}$$

where $\kappa(S) = \|S\|\,\|S^{-1}\|$ is the condition number of S and ρ and K are defined in (3.3).

As has been shown in [6], in general this number k can depend heavily on the initial vector and much better estimates can be derived than if we consider an arbitrary initial vector. However, in this presentation we consider (3.5) for a general vector e^0. In this case, we need only find the smallest integer k such that for some polynomial π_k^1 of degree k, normalized at the origin, we have

$$\min_{p_k \in \pi_k^1} \|p_k(J)\| \leq \frac{\rho\varepsilon}{K\kappa(S)} \ . \tag{3.7}$$

The problem of finding the minimal number of iterations has thus been reduced to a pure approximation problem. To proceed further, we assume that most of the eigenvalues can be found in an ellipse $E(a,b)$, symmetrically oriented along the real axis with foci a, b, and with $0 < a \leq b$ on this axis. In practice, this ellipse is chosen so that it contains most of the eigenvalues except some 'outliers', i.e., isolated eigenvalues outside the ellipse. Let q be the number of such 'outliers' and those eigenvalues in $E(a,b)$ which are deficient, i.e., which correspond to a pure Jordan block of size $(s_i > 1)$. Denote the eigenvalues by λ_i'. If there are two or more Jordan blocks belonging to a single eigenvalue, we take here the largest deficiency index. Then the following estimate holds.

Theorem 3.2 *[Axelsson, 1994; Axelsson, Makarov, 1995]. Assume that the eigenvalues of $B = C^{-1}A$ are located in the ellipse $E(a,b)$ except for some outliers as defined above. Then the smallest number k for which (3.5) holds is bounded by*

$$k \leq \sum_{i=1}^{q} s_i + \hat{k} \ ,$$

where

$$\hat{k} = \left[\ln\left(\frac{1}{\varepsilon'} + \sqrt{\frac{1}{(\varepsilon')^2} - 1} \right) \Big/ \ln \hat{\sigma}^{-1} \right]$$

Here,

$$\varepsilon' = \left[\max_{\lambda \in E(a,b)} \prod_{i=1}^{q} \left| 1 - \frac{\lambda}{\lambda_i'} \right|^{s_i} \right]^{-1} \frac{\varepsilon\rho}{K\kappa(S)},$$

$$\hat{\sigma} = \sigma\sqrt{\frac{1+\delta}{1-\delta}} \ , \qquad \sigma = (1 - \sqrt{a/b})/(1 + \sqrt{a/b}),$$

and δ is the ratio of the semi-axes of $E(a,b)$. When $\delta \ll 1$, and $\frac{a}{b} \ll 1$, it holds

$$1/\ln \hat{\sigma}^{-1} \sim \tfrac{1}{2}\sqrt{b/a} \ .$$

This theorem generalizes a similar theorem in [7] which dealt only with real eigenvalues.

This result shows that the number of iterations is bounded by the sum of the deficiency indices s_i and a number which depends essentially on the distribution of those eigenvalues which do not correspond to proper Jordan boxes. Note that for a normal matrix there are no deficient eigenvalues, so q contains only the outlier eigenvalues, and, furthermore, $\kappa(S) = 1$.

The theorem shows that outlier eigenvalues with a big absolute value compared to those in $E(a, b)$ can be annihilated with a single iteration while an outlier with a small absolute value causes a penalty in the form of a small additional factor in ε'. In the latter case one sees a delay (near stagnation) in the convergence before this eigenvalue has been annihilated. The number of additional iterations grows logarithmically with its inverse distance to the origin. When all outliers have been eliminated and when the ellipse is sufficiently far from the origin, one sees a superlinear rate of convergence. For further discussions on the above, see [7] and [5].

The estimate in Theorem 3.2 holds for any initial residual. A particular initial vector may be such that the initial residual decays rapidly (but sublinearly) during some initial iterations before the phase of linear convergence is entered. However, note that this does not necessarily mean a corresponding fast decay of the errors, in particular if the decay of the residuals was caused mainly be the annihilation of big eigenvalues. For a detailed discussion on the different convergence phases, frequently seen in practice, see [5], [6].

When B is highly non-normal with many Jordan boxes and/or Jordan boxes of high order or when $\kappa(S)$ is huge, the above estimate is less useful and there is an alternative approach to estimate the rate of convergence which is based on pseudo-eigenvalues (see [24]) which may give more accurate estimates in such cases.

As is well known, the information given by the eigenvalues is already insufficient in judging the convergence of a basic iteration method in the form

$$x^{k+1} = x^k - \alpha r^k.$$

The eigenvalues λ_i of $I - \alpha A$ may all have absolute values < 1 ($|\lambda_i| < 1$) so, asymptotically, the method converge as $(I - \alpha A)^k \to 0$, $k \to \infty$. However, in the initial transient phase $\|(I - \alpha A)^k\|$ can take huge values and in finite precision computation one may never enter the asymptotic phase.

An example of an inherently ill-conditioned matrix is the upper triangular Jordan block matrix J. Following [24], let J have order 32 with zero diagonal elements and let $J_{\widetilde{\varepsilon}}$ be the perturbation of J where the lower-left corner entry is perturbed to $\widetilde{\varepsilon} = 10^{-3}$. This small perturbation changes the zero eigenvalues of J (of multiplicity 32) into 32 distinct eigenvalues of magnitude $\widetilde{\varepsilon}^{\frac{1}{32}} \simeq 0.8$. This illustrates that the spectrum $\sigma(A)$ of a matrix is a bad measure of sensitivity of A w.r.t. perturbations. In fact, in general, the additivity property does not hold for the spectrum, i.e.

$$\sigma(A + E) \not\subset \sigma(A) + \sigma(E).$$

On the other hand, the numerical range, $W(x) = W(B,x) = \frac{(Bx,x)}{(x,x)}$, is additive since

$$W(A+E) \subset W(A) + W(E).$$

One can say that the spectrum is a too small set but, as it turns out, the numerical range is too large to give useful information. The pseudo eigenvalues could possibly be a good compromise between the two sets.

The pseudo-eigenvalues are defined as the set of eigenvalues of a perturbed matrix with perturbation satisfying $\|E\| \leq \widetilde{\varepsilon}$. (One can define a more general set of pseudo-eigenvalues by taking perturbations satisfying some other stability region than the disc with radius $\widetilde{\varepsilon}$. For this and other additional comments, see [16].)

Definition 3.1 Let $\Lambda_{\tilde{\varepsilon}} \supseteq \Lambda = \{\lambda_i\}$ denote the set of *$\tilde{\varepsilon}$-pseudo-eigenvalues* of A, i.e., all those points in the complex field, $z \in \mathbb{C}$ which are eigenvalues of some matrix $A+E$ with $\|E\| \leq \tilde{\varepsilon}$.

Now, if λ is an $\widetilde{\varepsilon}$-pseudo-eigenvalue of A, then there is a perturbation E, $\|E\| \leq \widetilde{\varepsilon}$, for which $(A+E)x = \lambda x$, $\|x\| = 1$ with eigenvector x. Thus $\|(\lambda I - A)x\| = \|Ex\| \leq \widetilde{\varepsilon}$. Further, for any matrix B, it is true that $\|B^{-1}\| \geq 1/\|Bx\|$, $\|x\| = 1$. Therefore it follows by letting $B = \lambda I - A$ that

$$\|(\lambda I - A)^{-1}\| \geq \frac{1}{\|(\lambda I - A)x\|} \geq \frac{1}{\widetilde{\varepsilon}}.$$

(Here the convention is that $\|(\lambda I - A)^{-1}\| = \infty$ when λ equals an eigenvalue of A.)

It can also be seen that the latter property implies that λ is an $\widetilde{\varepsilon}$-pseudo-eigenvalue. Therefore the following equivalences hold.

Theorem 3.3 *(Trefethen, 1992) Let A be a square matrix. Then the following are equivalent*

(i) *λ is an $\widetilde{\varepsilon}$-pseudo-eigenvalue.*

(ii) *$\|(\lambda I - A)x\| \leq \widetilde{\varepsilon}$ for some x, $\|x\| = 1$.*

(iii) *$\|(\lambda I - A)^{-1}\| \geq 1/\widetilde{\varepsilon}$.*

Note that $\|(\lambda I - A)^{-1}\|$ equals the inverse of the smallest singular value of $(\lambda I - A)^{-1}$. If A is normal then

$$\|(\lambda I - A)^{-1}\| = \frac{1}{\mathrm{dist}(\lambda, \sigma(A))},$$

i.e., equals the inverse of the distance from λ to the spectrum of A, implying a tentlike shape hanging from its poles. In the non-normal case, the shape of the surface can be much more complicated and $\|(\lambda I - A)^{-1}\|$ can attain huge values even when λ is far from an eigenvalue.

Let $L(\tilde{\varepsilon})$ be the arclength of the boundary $\Gamma(\tilde{\varepsilon})$ of $\Lambda_{\tilde{\varepsilon}}$. It is known that, for any polynomial p_k, the matrix $p_k(A)$ can be written as a Cauchy integral,

$$p_k(A) = \frac{1}{2\pi i} \int_{\Gamma(\tilde{\varepsilon})} p_k(z)(zI - A)^{-1}\, dz\ , \tag{3.8}$$

when $\Gamma(\tilde{\varepsilon})$ is any simple closed curve or the union of several simple closed curves containing the spectrum of A. Considering a contour $\Gamma(\tilde{\varepsilon})$ on which $\|(zI - A)^{-1}\| = \tilde{\varepsilon}^{-1}$ and taking norms in (3.8), we obtain

$$\|p_k(A)\| \leq \frac{L(\tilde{\varepsilon})}{2\pi\tilde{\varepsilon}} \max_{z \in \Gamma_{\tilde{\varepsilon}}} |p_k(z)| \ . \tag{3.9}$$

Large values of $L(\tilde{\varepsilon})/\tilde{\varepsilon}$ indicate that A is highly sensitive to perturbations of its coefficients. This sensitivity will be reflected in severe problems when solving systems with A or when computing eigenvalues of A. To make the estimate (3.9) viable, one must choose $\tilde{\varepsilon}$ properly. Choosing $\tilde{\varepsilon}$ small gives too large values of $L(\tilde{\varepsilon})/\tilde{\varepsilon}$ while choosing $\tilde{\varepsilon}$ too large may increase the set $\Lambda_{\tilde{\varepsilon}}$ and, hence $L(\tilde{\varepsilon})$ too much.

Theorem 3.4 *Let $L(\tilde{\varepsilon})$ and $\Gamma(\tilde{\varepsilon})$ be defined as above. Then the residuals and iteration errors in the* GCG-MR *and the* GCG-OR *methods satisfy*

$$\begin{aligned} \frac{\|r^k\|}{\|r^0\|} &\leq \frac{L(\tilde{\varepsilon})}{2\pi\tilde{\varepsilon}} \min_{p_k \in \pi_k^1} \max_{z \in \Gamma(\tilde{\varepsilon})} |p_k(z)|, && \textit{in the } \text{GCG-MR } \textit{method} \\ \frac{\|e^k\|}{\|e^0\|} &\leq \frac{K}{\rho} \frac{L(\tilde{\varepsilon})}{2\pi\tilde{\varepsilon}} \min_{p_k \in \pi_k^1} \max_{z \in \Gamma(\tilde{\varepsilon})} |p_k(z)|, && \textit{in the } \text{GCG-OR } \textit{method.} \end{aligned}$$

In both estimates in Theorem 3.4 we have avoided the appearance of the, potentially huge, factor $\kappa(S)$. For certain problems, in particular when A is far from normal, and with a proper choice of $\tilde{\varepsilon}$, the bound in Theorem 3.4 may be smaller than the bound in Theorem 3.2. In such cases the convergence of the GCG-MR and GCG-OR methods depend on a polynomial approximation problem defined on a pseudo-spectrum rather than on just the spectrum. As it turns out, neither bound is sharp in general. (On the other hand, it can be readily seen that the bound for the GCG-MR method,

$$\frac{\|r^k\|}{\|r^0\|} \leq \min_{p_k \in \pi_k^1} \max_{z \in \Lambda(A)} |p_k(z)| \tag{3.10}$$

holds for normal matrices and is sharp in the respect that for any k there is an initial residual for which equality holds in (3.10)).

4 Concluding remarks

It has been shown in Theorem 3.1 that the generalized conjugate gradient (GCE) or minimal residual method can have excellent convergence properties. However, the actual convergence depends much on the distribution of eigenvalues or of pseudoeigenvalues.

Unless truncated, the computational cost per iteration step of the method grows linearly with the iteration number but this is avoided when restarts are used. In a companion paper to the present one it will be shown that the GCG method can be implemented as efficiently as the GMRES method but it has the advantage over the latter that the residual is available at any step and therefore one can stop the iterations as soon as the residual becomes sufficiently small. In the GCG method it is

also possible to use other vectors than Krylov vectors and combinations of Krylov set vectors and certain approximate solution vectors (see [21]) can often be particularly efficient.

In practice, even after relatively few iterations, rounding errors cause a deviation of the convergence from an ideal method with no rounding errors. This topic is treated in particular details in [14]. The exact orthogonality among vectors is lost and this causes an increase of the number of iterations in the GMRES and, in particular, in methods using short term recursions, as all short term recursions depend upon certain orthogonality properties. The GCG method is less dependent on orthogonality as it is based on a minimization over the current vector set. It should, however, be implemented with restart when the convergence tends to stagnate. The use of a new set of Krylov vectors based on a new preconditioner has thereby turned out to be particularly efficient, see [9].

For a nice survey of various phenomenae which can occur in other methods due to loss of orthogonality and also the dependence of the rate of convergence on various eigenvalue distributions, see [11]. See also [15].

In the companion paper to the present one it will also be shown that the orthogonal residual method can have very large residuals, which occur when there is stagnation in the convergence of the corresponding minimum residual method.

Acknowledgement

Helpful comments on an earlier version of the paper and also several valuable discussions on related topics during the past several years with David M. Young are kindly acknowledged.

References

[1] W.E. Arnoldi, The principle of minimized iterations in the solution of the matrix eigenvalue problem, *Quart. Appl. Math.* 9 (1951), 17-29.

[2] S.F. Ashby, T.A. Manteuffel, and P.E. Saylor, "A taxonomy for conjugate gradient methods", *SIAM J. Numer. Anal.*, 27 (1990), 1542-1568.

[3] O. Axelsson, "Conjugate gradient type methods for unsymmetric and inconsistent systems of linear equations", *Linear Algebra and its Applications* 29 (1980), 1-16.

[4] O. Axelsson, "A generalized conjugate gradient, least squares method", *Numer. Math.* 51 (1987), 209-227.

[5] O. Axelsson, *Iterative Solution Methods*, Cambridge Univ. Press, New York, 1994.

[6] O. Axelsson, "Condition numbers for the study of the rate of convergence of the conjugate gradient method", *Iterative Methods in Linear Algebra* II, IMACS, Blagoevgrad 1995, pp. 3-33.

[7] O. Axelsson, G. Lindskog, "On the rate of convergence of the conjugate gradient method", *Numer. Math.* 48(1986), 499-523.

[8] O. Axelsson and M. Makarov, "On a generalized conjugate gradient orthogonal residual method", *Numerical Linear Algebra with Applications*, 2(1995), 467-480.

[9] O. Axelsson and M. Nikolova, "A generalized conjugate gradient minimum residual method (GCG-MR) with variable preconditioners and a relation between residuals of the GCG-MR and GCG-OR methods", *Communications in Applied Mathematics* 1 (1997), 371-388.

[10] C.G. Broyden, A new taxonomy of conjugate gradient methods, *Computers and Mathematics with Applications*, 31(4/5), 7-17, 1996.

[11] T.A. Driscoll, K.-C. Toh, L.N. Trefethen, From potential theory to matrix iterations in six steps, *SIAM Rev.* 40 (1998), 547-578.

[12] V. Faber and T. Manteuffel, "Necessary and sufficient conditions for the existence of a conjugate gradient method", *SIAM J. Numer. Anal.* 21(1984), 352-367.

[13] R. Fletcher, "Conjugate Gradient Methods for Indefinite Systems". *Proceedings of the Dundee Conference on Numerical Analysis*, 1975, *Lecture Notes in Mathematics* no. 506, pages 73-89, Springer-Verlag, New York, 1976.

[14] A. Greenbaum, Iterative Methods for Solving Linear Systems, *SIAM*, Philadelphia, 1997.

[15] A. Greenbaum, V. Ptàk and Z. Strakoš, Any nonincreasing convergence curve is possible for GMRES, *SIAM J. Matrix Anal. Appl.* 17 (1996), 465-469.

[16] K.E. Gustafson and D.K.M. Rao, *Numerical Range, The Field of Values of Linear Operators and Matrices*, Springer, 1997.

[17] L.A. Hageman and D.M. Young, *Applied Iterative Methods*, Academic Press, New York, 1981.

[18] M.R. Hestenes and E. Stiefel, "Methods of Conjugate Gradient for Solving Linear Systems", *Journal of Research of the National Bureau of Standards*, 49(6) , 409-436, Dec. 1952.

[19] K.C. Jea and D.M. Young, "The simplification of generalized conjugate gradient methods for nonsymmetrizable linear systems", *Linear Algebra and Its Applications*, 52/53, 399-417, 1983.

[20] W.D. Joubert and D.M. Young, "Necessary and sufficient conditions for the simplification of generalized conjugate gradient algorithms", *Linear Algebra and its Applications*, 88/89 (1987), 449-485.

[21] I.E. Kaporin and O. Axelsson, On a class of nonlinear equation solvers based on the residual norm reduction over a sequence of affine subspaces, *SIAM J. Sci. Comput.*, 16 (1995), pp. 228-249.

[22] Y. Saad and M.H. Schultz, "Conjugate gradient-like algorithms for solving nonsymmetric linear systems", *Math. Comput.* 44(1985), 417-424.

[23] Y. Saad and M.H. Schultz, "GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems", *SIAM J. Sci. Stat. Comput.*, 7(1986), 856-869.

[24] L.F. Trefethen, Pseudospectra of matrices, *in* Numerical Analysis 1991, D.F. Griffiths and G.A. Watson, eds., Longman Scientific and Technical, Harlow, UK, 1992, pp. 234-266.

[25] P.K.W. Vinsome, "ORTHOMIN, an Iterative Method for Solving Sparse Sets of Simultaneous Linear Equations", *4th Symposium of Numerical Simulation of Reservoir Performance, the Society of Petroleum Engineers of the AIME*, Paper SPE 5739, 1976.

[26] H. van der Vorst, "BI-CGSTAB: a fast smoothly converging variant of BI-CG for the solution of nonsymmetric linear systems", *SIAM J. Sci. Stat. Comput.*, 13(1992), 631-644.

[27] D. Young and K. Jea, "Generalized conjugate gradient acceleration of nonsymmetrizable iterative methods", *Linear Algebra and its Appl.*, 34(1980), 159-194.

[28] R. Weiss, *Parameter-Free Iterative Linear Solvers*, Akademie Verlag, Berlin 1996.